Re-examining how utility and weighting functions get their shapes: a quasi-adversarial collaboration providing a new interpretation.

Despoina Alempaki[1], Emina Canic[2], Timothy L. Mullett[1], William J. Skylark[3], Chris Starmer[4], Neil Stewart[1], Fabio Tufano[4]

[1]Warwick Business School, University of Warwick
[2] Department of Psychology, University of Warwick
[3] Department of Psychology, University of Cambridge
[4] School of Economics, University of Nottingham

Abstract

Stewart, Reimers and Harris (2015, *SRH* hereafter) demonstrated that shapes of utility and probability weighting functions could be manipulated by adjusting the distributions of outcomes and probabilities on offer, as predicted by the theory of Decision by Sampling. So marked were these effects that, at face value, they profoundly challenge standard interpretations of preference theoretic models where such functions are supposed to reflect stable properties of individual risk preferences. Motivated by this challenge, we report an extensive replication exercise based on a series of experiments conducted as a quasi-adversarial collaboration across different labs and involving researchers from both economics and psychology. We replicate the SRH effect across multiple experiments involving changes in many design features; importantly, however, we find that the effect is also present in designs modified so that Decision by Sampling predicts no effect. While those results depend on model-based inferences, an alternative analysis using a model free comparison approach finds no evidence of patterns akin to the SRH effect. On the basis of simulation exercises, we demonstrate that the SRH effect may be a consequence of misspecification biases arising in parameter recovery exercises that fit imperfectly specified choice models to experimental data. Overall, our analysis casts the SRH effect in an entirely new light.

Keywords: utility, probability weighting, replication, Decision by Sampling, risky choice

Acknoweldgements: This work was supported by the Economic and Social Research Council [grant numbers ES/N018192/1, ES/K002201/1, and ES/P008976/1] and the Leverhulme Trust [grant RP2012-V022]. We thank the editor, an associate editor and anonymous referees for very helpful comments.

Introduction

In a recent paper published in this journal, Stewart, Reimers and Harris (2015) presented evidence from a series of experiments putatively demonstrating that the utility and probability weighting functions revealed by fitting standard economic models to binary choice data were highly sensitive to changes in the distributions of payoffs and probabilities in the choice sets. While the existence of some such sensitivity may be no surprise (e.g., Drichoutis and Nayga, 2013; Etchart-Vincent, 2004; Fehr-Duda et al. 2010, 2011), the extent of malleability identified by SRH is considerable. For example, for some distributions of probabilities and payoffs, SRH were able to produce concave utility functions and inverse-S shaped probability weighting functions as commonly reported elsewhere in the literature; yet, for other distributions they generated mirror image patterns (i.e., convex utility and S-shaped probability weighting functions). For convenience, we will refer to the apparent malleability of the utility and probability weighting functions identified by SRH as the *SRH effect*.

At face value, the SRH effect provides important new support for the model of Decision by Sampling because - as we explain in Section II – predictions of this model (which we refer to as DbS for short) prompted its discovery. More broadly, however, the SRH effect sets a potentially severe challenge to a wide range of models of risky decision making in the preference-theoretic tradition which interpret utility and weighting functions as embodying an individual's risk preference. If a researcher can, as SRH explicitly suggest, choose the shapes of the functions they wish to reveal by adjusting the set of gambles used to elicit them, then the interpretation that such procedures reveal underlying preferences is undermined. Hence, the SRH effect provides seemingly powerful new ammunition for those critical of the adequacy of preference-based models of risky-choice (Friedman et al. 2014; Gigerenzer, 2016) and support to those who favour process based models, and in particular, the model of Decision by Sampling (Stewart, 2009; Stewart et al. 2006).

But before interpreting the SRH effect as a strong challenge to preference based models (or support for procedural models including DbS), it is appropriate to question whether the effect is replicable and robust. That question is pertinent, not least, in the light of contemporary controversy surrounding the replicability of many of the findings in the behavioural sciences and elsewhere (e.g., Camerer et al. 2016; Maniadis et al. 2014; Open Science Collaboration, 2015). Given this background controversy and the challenging nature of the SRH findings, we believe that good scientific practice demands careful scrutiny of the SRH effect, via attempts at replication, to properly assess its significance. With this motivation in mind, we report an extensive set of replication experiments investigating two key issues: First, we examine whether the SRH effect is replicable and robust to variations in experimental design. Pre-empting our results, we conclude that the SRH effect is replicable, and robust to many small variations in experimental design. The second key issue concerns

the origins of the effect. Three dimensions of our results point to a different interpretation of the SRH effect. First, we are able to reproduce the SRH effect in designs which turn off the mechanism that, according to DbS, generates it; Second, choice based tests find no evidence of the effects predicted by DbS; Third, by exploring parameter recovery in simulations of a stochastic expected utility model, we find evidence that biases due to model misspecification may plausibly explain the SRH effect.

In what follows, we report a set of 14 new experiments conducted as part of what we call a *quasi-adversarial collaboration* and combine these with a reanalysis of the 5 original experiments. The term "adversarial collaboration" has been used to refer to experimental research projects jointly planned and executed by two or more researchers (or research groups) who have ex-ante conflicting hypotheses about its outcome (for discussion and examples see Bateman et al. 2005; Corrigan, 2011; Kahneman, 2003; Latham et al. 1988; Mellers et al. 2001). While our collaboration does not have exactly this form (hence the qualifier 'quasi'), the seven researchers involved in this collaboration come from different disciplines (economics and psychology), different labs, and have very different degrees of prior investment in the competing theoretical frameworks that would be supported or challenged by the existence of the SRH effect. We also use the qualifier "quasi' to signal that the set of experiments reported here did not emerge from a common plan of adversarial collaboration agreed before any of the experiments began. Instead, our collaboration emerged as subsets of the present co-authors began to discover that we were undertaking very closely related work exploring the SRH effect, independently, at different labs. We then began to compare results and, later, to discuss designs for new experiments. Further experiments were subsequently run by different sub-groups of us, based in three different labs at two universities, using a mixture of lab-based and online protocols. The development of the designs involved varying degrees of consultation between us, as well as key variations in designs and procedures, which we document below. Through this process we have generated a rich source of evidence relative to the SRH effect, which we bring together in this paper.

The somewhat organic evolution of the collaboration does not mean that the set of replications, when viewed as a whole, lacks structure. We will argue that, although we did not set out with this explicit purpose, the resultant set of experiments reflect and, indeed, extend a replication strategy proposed by Levitt and List (2009). They advocate a methodology involving replication at three levels: reanalysing data from the original study to be replicated; running fresh experiments using designs approximating the experiment to be replicated and thirdly, conditional on replicating the original results, running experiments to probe origins of the phenomenon observed. Our experiments involve replication at all three levels but we also add a further dimension to our analysis. Through our experiments, we generated a rich data set based on decisions of 1880 subjects which we use to run a composite analysis combining

SRH's original data with data from our new experiments. We refer to this below as our 'meta-analysis'[1]. This complements the individual experiments by placing confidence bounds on the size of the SRH effect and allowing assessment of how it varies with some key design features of our replications. As such we interpret part of our contribution as piloting an extended, four level, version of the Levitt and List (2009) methodology enhanced by meta-analysis. While the meta-analysis confirms a non-zero effect size, our experiments and simulations combined, cast the SRH effect in an entirely new light.

The paper is organized as follows. Section II reviews key features and findings of the original SRH study. Section III presents results from all four levels of the replication process. Section IV considers alternative explanations for the SRH effect and Section V concludes.

I. The Original SRH (2015) study: setup, motivation, methodology and results

The main results in the original SRH study are based on data generated from experiments in which individuals had to make a series of choices between pairs of gambles of the form "p chance of x, otherwise nothing" or "q chance of y, otherwise nothing" (where $p<q$ and $y<x$). SRH used these data to estimate utility functions over monetary payoffs (x, y) by fitting an expected utility model and probability weighting functions over probabilities (p, q) by fitting a subjective expected utility model.

All experiments followed the same logic whereby, for any given treatment, a fixed set of (five or six) money amounts was fully crossed with a fixed set of (five or six) probabilities to create a set of gambles. This was then used to generate a set of pairwise choices, for any given treatment of the experiment, comprising all possible non-identical and stochastically non-dominant pairwise choices from the full set of gambles. A further 30 choices were added in which one option stochastically dominated as a catch for participants paying insufficient attention; those who chose dominated options in more than 10% of catch trials were excluded from the main analysis. The order of choices was randomized across participants.

There were two treatments in each experiment which varied according to either the skew in the distribution of money amounts, or the skew in the distribution of probabilities, used in construction of the choice sets. Each experiment then involved comparison of a treatment with a positive-skew distribution (of money amounts or probabilities) against a treatment with either a negative-skew or a zero-skew distribution (of money amounts or probabilities). It was comparisons between these pairs of treatments which generated the SRH effect. The actual amounts and probabilities used in five of the different experiments reported by SRH are depicted in Table 1.

[1] Although this analysis exploits all the currently existing data that we are aware of relating to the SRH effect, it is a limited exercise in the sense of relying on data that come from our own and from SRH's previous experiments.

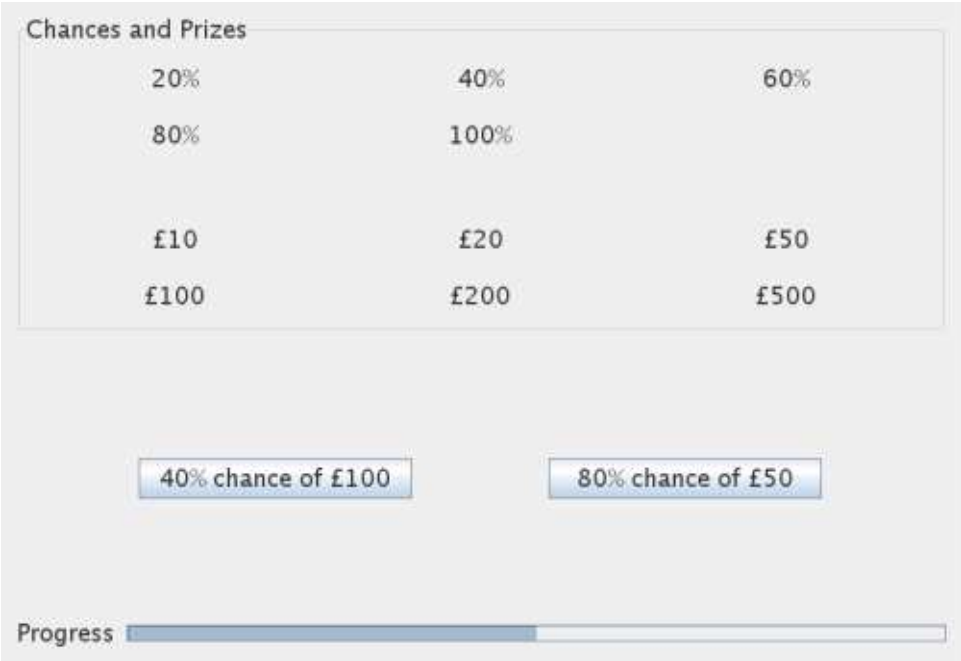
Table 1. Amounts and probabilities used in the original SRH experiments to create the choice sets.

Experiment*	Domain	Skew	Amounts (£)	Probabilities (%)
SRH 1A	Utility	Positive vs. Negative	10, 20, 50,100, 200, 500 vs. 10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100
SRH 1B	Utility	Positive vs. Zero	10, 20, 50,100, 200, 500 vs. 0, 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100
SRH 1C	Utility	Positive vs. Zero	10, 20, 50,100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100
SRH 2A	Probability	Positive vs. Negative	100, 200, 300, 400, 500 vs. 100, 200, 300, 400, 500	10, 20, 30, 40, 70, 90 vs. 10, 30, 60, 70, 80, 90
SRH 2B	Probability	Positive vs. Negative	100, 200, 300, 400, 500 vs. 100, 200, 300, 400, 500	1, 2, 5, 10, 50, 99 vs. 1, 50, 90, 95, 98, 99

**Note:* There was an additional series of temporal discounting experiments, which we do not address in this paper.

Notice that the first three experiments in Table 1 (Experiments SRH 1A-1C) test for the SRH effect in the utility domain by varying the distributions of the money amounts between treatments, holding constant the set of probabilities used to construct the set of gambles. For example, in Experiment SRH 1A, the gambles are constructed using a common set of probabilities in both treatments (ranging from 20% to 100% in 20% steps); whereas the money amounts range from £10 to £500 in both treatments but for one treatment (the *positive-skew* distribution) the intermediate outcomes are all in the lower half of the range, while for the other treatment (the *negative-skew* distribution) all of the intermediate outcomes are in the upper half of the range. These experiments test how changing the distribution of money amounts changes the revealed utility functions. In the last two experiments depicted in Table 1 (SRH 2A and 2B) the distribution of amounts is common across treatments for each experiment, but the distribution of probabilities changes between them. These experiments tested for the SRH effect in the probability domain by examining the sensitivity of resulting probability weighting functions to changes in the distribution of probabilities.

Figure 1. Interface used in SRH 1A.



An example of the choice interface, based on SRH 1A, is shown in Figure 1. At the top of the screen participants saw the distributions of chances to win and prizes on offer across the set of choices. In most experiments, this information was on screen continuously while subjects made their decisions.[2] Although subjects were not explicitly informed about

[2] Having this information on screen continuously does not appear to be a decisive factor. For Experiment SRH 1B, subjects completed the series of choices without the distribution of chances and prizes on offer being shown on

the number of choices they had to make they were told about the likely duration of the experiment and a bar at the bottom of the screen kept track of their progress. The number of unique non-dominated pairwise choices was 150 for Experiments SRH 1A and SRH 2B, 120 for Experiments SRH 1B and SRH 2A and 100 for Experiment SRH 1C. After making each decision, the next choice appeared automatically. In any given choice, the two gambles were presented in the form of text on separate 'buttons' and subjects indicated their decision by clicking on one of them. They were told that at the end of the experiment one of their choices would be randomly selected and their chosen gamble would be played out and paid for real using an exchange rate: 1 pound equals 1 pence.[3] All SRH original experiments were conducted at the University of Warwick (with one run online).

In a moment we will review the main findings of SRH. As a prelude to that, we note that the treatment comparisons in SRH have a particular theoretical motivation because, as SRH explain, the DbS model predicts systematic differences between them. The DbS model is a mechanism for the construction of choices from a series of ordinal comparisons between pairs of attribute values. Readers interested in the details of the DbS model should consult Stewart (2009), Stewart et al. (2006), and SRH (2015). For now, the following property is sufficient: Because the probability that an attribute value will win an ordinal comparison is given by its rank position within those attribute values available, DbS predicts people will choose as if subjective value is the *rank position* of an attribute value within all those available. For example, consider the evaluation of the amount £200 in the context of the distributions used in Experiment SRH 1B. In the positive-skew treatment where the amounts are £10, £20, £50, £100, £200, and £500, the £200 outcome is better than 4 out of 5 of the other outcomes that will be encountered. But now consider the evaluation of £200 in the context of the zero-skew treatment of Experiment SRH 1B with the distribution of £0, £100, £200, £300, £400, £500. In this case, £200 is better than only 2 out of 5 other outcomes that will be encountered. Thus, DbS implies a *higher* utility for £200 in the positive-skew treatment, compared to its utility in the zero-skew treatment.

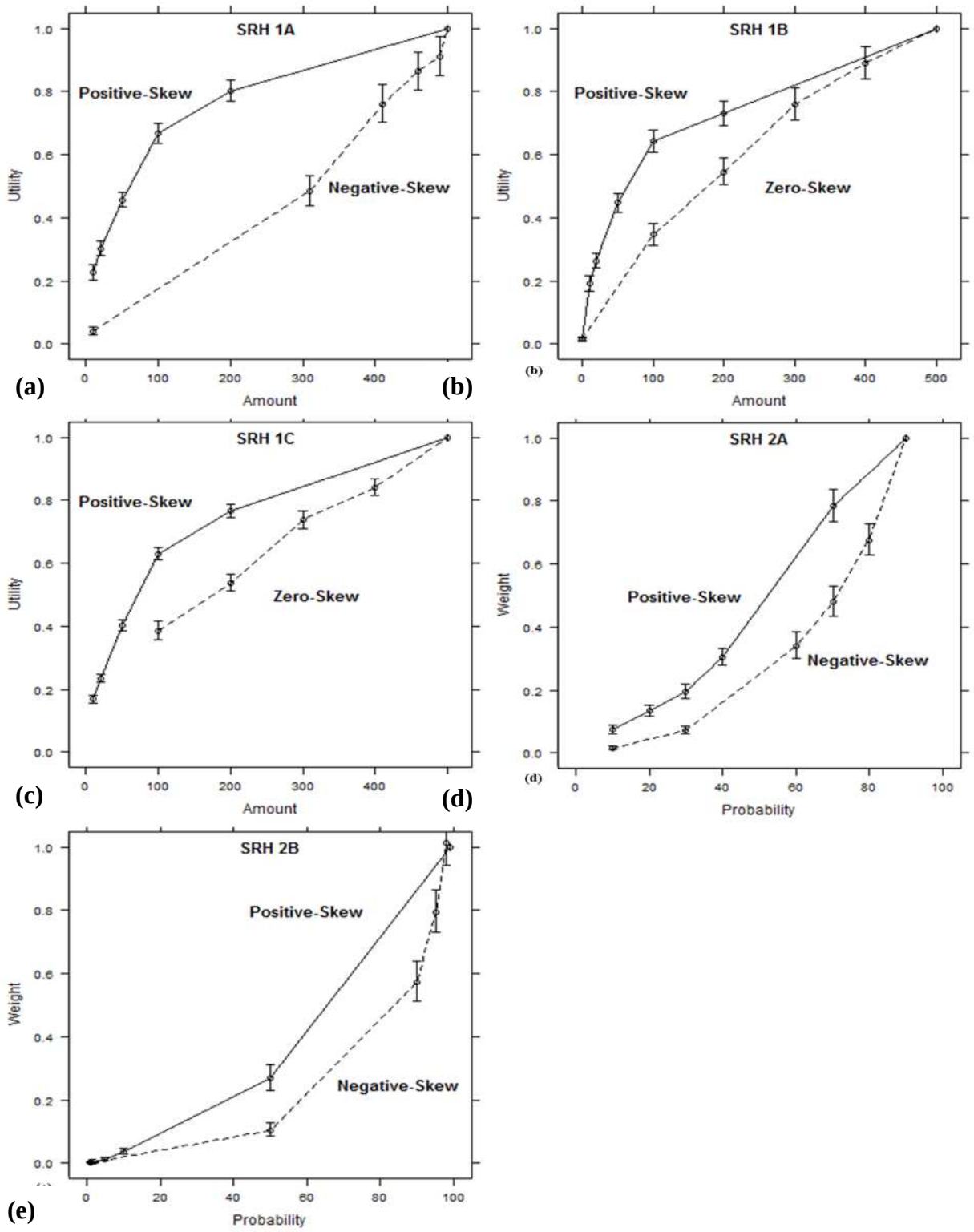
Figure 2 shows the estimated utility and weighting functions for all five SRH original experiments. In line with DbS predictions, the utility functions (utility experiments, Figures 2a, 2b, 2c) and the weighting functions (probability experiments, Figures 2d, 2e) were more concave when the skew in the distribution of amounts or probabilities was positive than when it was negative or zero. For example, the utility for £200 in SRH 1B is lower in the zero-skew treatment than in the positive-skew treatment. The intuition from DbS is that, because

screen during choices. This experiment produced qualitatively very similar results to Experiment SRH 1C which used almost identical amounts and did have the distribution of chances and prizes on screen.

[3] In Experiment SRH 1A two choices were randomly selected for payment. The exchange rate was halved for this experiment. In Experiment SRH 1C the choices were hypothetical rather than incentivized.

subjective value is given by rank position, subjective value must increase most quickly where attribute value densities are highest. This means a steeper increase early on, for distributions with positive-skew compared to distributions with negative- or zero-skew.

Figure 2. The revealed utility and probability weighting functions from SRH.



Note: Error bars are 95% confidence intervals.
Source: Adapted from SRH (Figures 4a, 4b, 4c, 5a, and 5b respectively).

II. Four levels of replication

We now present the results of our new analysis. This involves four levels of replication, extending the methodology advocated by Levitt and List (2009) with an additional stage of meta-analysis to take advantage of the rich data set generated via our new experiments. In terms of data analysis, our strategy is to use methods which are essentially the same as those applied by SRH, except for some refinements explained below. We then hold constant the statistical methods that we apply across the four levels of replication reported in this study.[4] Analysis for each of the four levels is presented as a separate subsection.

A. Level 1: Replication by reanalysing the original data

As step one, we reanalysed SRH's original experimental data to estimate the revealed utility and weighting functions. Specifically, we fit an expected utility model with a Luce (1959)-Shepard (1957)[5] choice rule incorporating a stochastic component to estimate utilities.

$$Prob(Choose\ safe) = \frac{bias\ (q\ u(y))^{\gamma}}{bias\ (q\ u(y))^{\gamma}+(p\ u(x))^{\gamma}} \quad (1)$$

Here $u(y)$ is the utility of the safer gamble's prize which occurs with probability q, while $u(x)$ is the utility of the riskier gamble's prize which occurs with probability p; $bias$ is a general tendency to choose safe irrespective of the actual amounts and probabilities on offer; and γ controls the level of determinism in responding (γ =1 gives choice probabilities proportional to the expected utilities, and γ >1 gives more extreme choice probabilities, so gambles with a slightly higher expected utility are very likely to be chosen).

The advantage of this model, though it is perhaps not obvious, is that for simple gambles it can be estimated as the following logistic regression:

$$\log\left[\frac{Prob(Choose\ safe)}{1-Prob(Choose\ safe)}\right] = \nu + \omega \log\left(\frac{q}{p}\right) + \sum_i \beta_i\ X_i \quad (2)$$

In Equation 2, each X_i is a dummy variable indicating the presence of amount *i* as *y* (coded +1), as *x* (coded -1), or absent from the choice (coded 0); $\nu = \log(bias)$; $\omega = \gamma$; and the utility of each amount $u(i) = \exp(\beta_i/\gamma)$.

A corresponding logistic regression for estimating weights for probabilities is obtained by exchanging the roles of *p* and *q* and of *x* and *y*.

$$\log\left[\frac{Prob(safe)}{1-Prob(safe)}\right] = \nu + \omega \log\left(\frac{y}{x}\right) + \sum_i \beta_i\ X_i \quad (3)$$

where the weighting of each probability is given by $w(p_i) = exp(\beta_i/\gamma)$.

SRH give further details in their Appendix A.

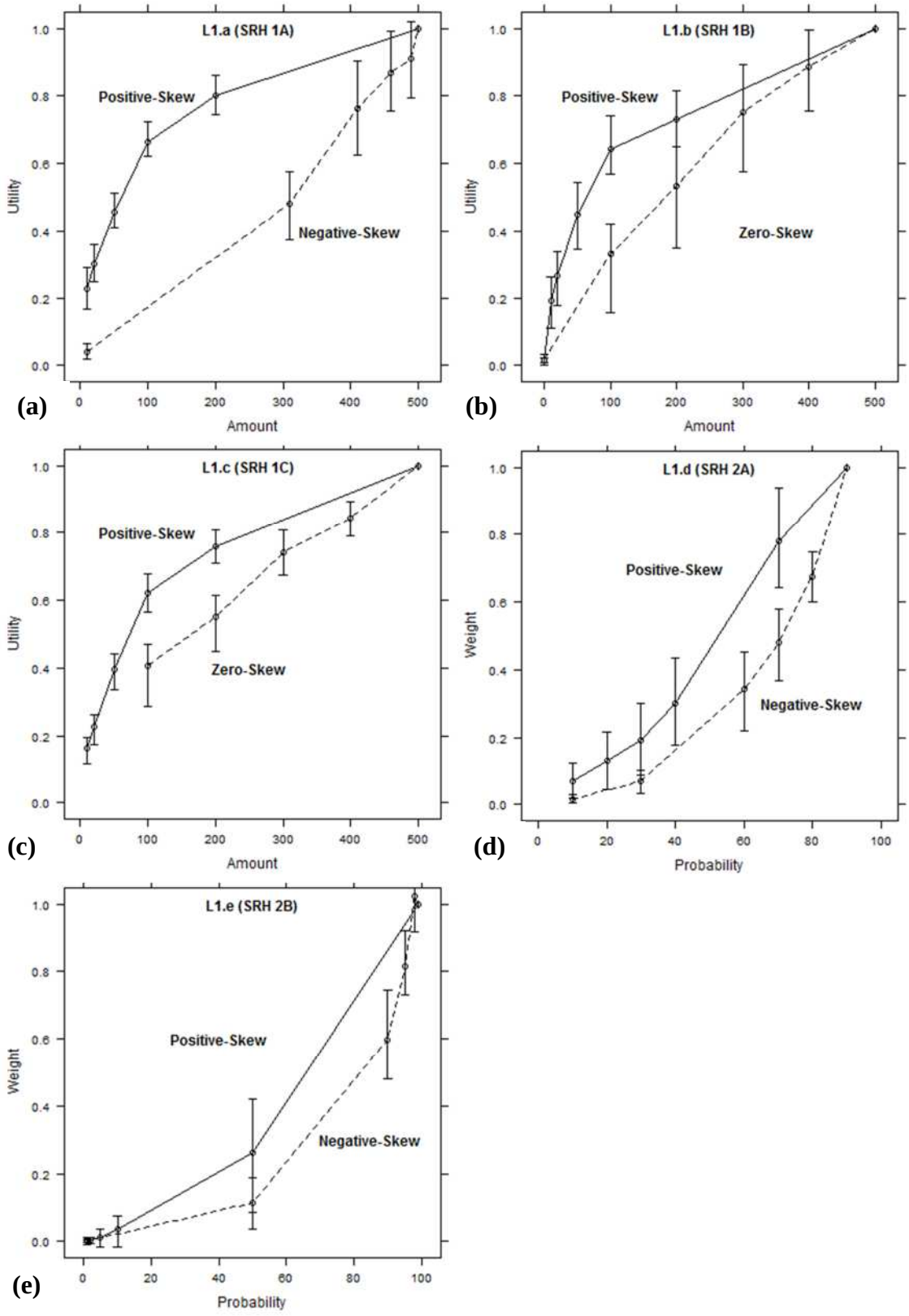
[4] While one could entertain different approaches to modelling, basing our approach on that used by SRH minimizes the chance that differences between our results and theirs are due to modelling differences; and holding the approach constant within our analysis allows us to rule out the possibility that differences across our experiments or levels of replication could be plausibly attributed to our statistical modelling methodology.

[5] Our results are qualitatively the same using the multinomial logit model of McFadden (1976, 2001).

Our analysis departs from the original analysis in SRH by using a different approach to estimate confidence intervals. We used a more reliable bootstrapping method instead of the standard errors from the model fits used by SRH, because we wanted to allow for the possibility of asymmetric confidence intervals. Furthermore, we estimated the revealed functions separately for each treatment. SRH used one model for both treatments, but this is not ideal; random effects cannot be estimated for all amounts for all participants, because each participant experienced only a subset of amounts. While Level 1 replication identified some minor calculation errors in SRH's original analysis, these did not change the conclusions of the original paper. A detailed comparison between SRH's original analysis with and without those calculation errors can be found in Appendix A.

The revealed functions from the replication analysis are shown in Figure 3. The panels correspond to those in Figure 2, where the original experiments are presented. For example, a direct comparison for Experiment SRH 1A shows that the replicated functions depicted in Figure 3a strongly mirror the original functions depicted in Figure 2a. The same result holds for all experiments. Hence, we conclude that the Level 1 analysis successfully replicates the SRH effect in both the utility and probability domains.

Figure 3. The revealed functions obtained from the replication analysis.



Note: Error bars are 95% confidence intervals.

B. Level 2 Replication using variants of the original design

In this section, we report a series of eight new experiments run by different subsets of the current authors. Each was designed to replicate one of the original SRH experiments but with several design variations introduced across our studies as set out below.

Table 2 presents the details of the Level 2 experiments. For example, at the top of the table, the experiment labelled "L2.a" is a replication of the original Experiment SRH 1C comparing positive-skew and zero-skew distributions of money amounts. The distributions of money amounts and probabilities correspond exactly with those in SRH 1C. Subsequent columns of Table 2 indicate that L2.a was an incentivized experiment with 54 participants at University 1 using 75 randomly selected choices from the original study.

Looking further down Table 2, you will see that we have replicated examples of both the utility and the probability weighting experiments, though there is a focus on the utility domain. This is partly an accident of history reflecting decisions made in the different labs when they started running these experiments independently. But, a focus on the utility domain may, nevertheless, be useful for several reasons. First, utility is arguably a more fundamental concept, compared to probability weighting, in models derived from the preference theoretic framework; it features in a wider class of models and it is the core subjective dimension in what has been the leading theory of risk preference–that is, expected utility theory.

While this focus seems justified by these arguments, we also wished to include at least one Level 2 test in the probability domain – hence the inclusion L2.h (additional experiments with manipulations of probability are reported as part of Level 3 replications).

Notice that across the series of experiments there is variation in: the skew (positive- vs negative- or zero-skew), the domain (utility or probability), the location (group that conducted the experiment), whether the experiment was conducted online or in the lab, the number of participants, the number of trials and the incentives (by using both incentivized and hypothetical experiments).

We think there is some advantage in focusing on particular SRH experiments to see the effects of small changes in procedures holding constant the distributions of amounts and probabilities. For this purpose we used SRH 1C, which contrasted positive-skew with zero-skew distributions of outcome values making Experiments L2.a-L2.e different replications of SRH 1C. The advantages of choosing to replicate mainly the Experiment SRH 1C are twofold: First, we can use the difference between the utilities of the common amounts for a direct comparison (there are no common amounts between the positive-negative conditions); Second, the round amounts used in the zero-skew condition are more representative of amounts often experienced by subjects in other experiments.

Table 2. Outline of the properties of all replication experiments from Level 2.

Replication (Original)	Skew	Domain	Amounts (£ or $)*	Probabilities (%)	Location (Sample)	N‡	No. Trials¥	Incentives
L2.a (SRH 1C)	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (Student sample)	54	Mean of 75 randomly selected	Course credit plus up to £5
L2.b (SRH 1C)	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (Prolific Academic)	200	150	£1.80 (non-incentivized)
L2.c (SRH 1C)	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (MTurk)	492	40	$1.80 (non-incentivized)
L2.d (SRH 1C)	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (MTurk)	145	Mean of 75 randomly selected	$1.80 (non-incentivized)
L2.e (SRH 1C)	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (50% MTurk and 50% Prolific Academic)	183	150	$2.25 or £1.50 and up to $/£25
L2.f (SRH 1A)	Positive vs. Negative	Utility	10, 20, 50, 100, 200, 500 vs.10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 2 (Student sample)	40	180	£0 up to £5
L2.g (New distribution)	Positive vs. Negative	Utility	5, 10, 20, 50, 100, 200, 500 vs. 5, 300, 400, 450, 480, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (MTurk)	154	180	$3 (non-incentivized)
L2.h (SRH 2B)	Positive vs. Negative	Probability	100, 200, 300, 400, 500 vs. 100, 200, 300, 400, 500	1, 2, 5, 10, 50, 99 vs. 1, 50, 90, 95, 98, 99	University 2 (Student sample)	29	180	£0 up to £5

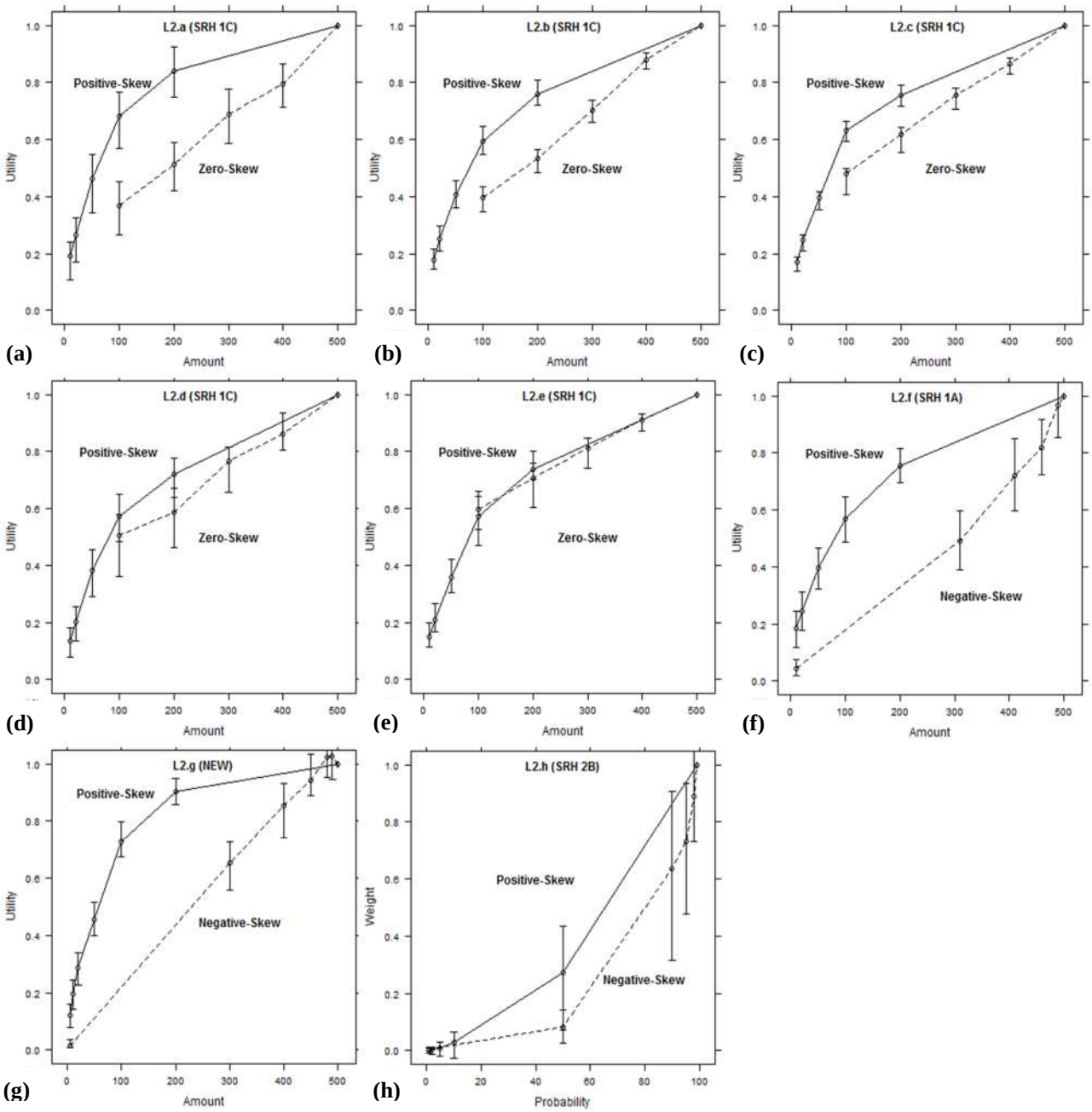
Notes: *We used $ for all Amazon Mechanical Turk (MTurk) samples and £ for all Prolific Academic and student samples. In Experiments L2.f and L2.h, where the experiment was conducted using z-tree (Fischbacher, 2007) and subjects were recruited via the online system ORSEE (Greiner, 2015), a show-up fee of £7 was added to the earnings from the experiment.

‡ Catch trials were not included in replication Experiments L2.b, L2.c, L2.d, L2.e and L2.g. 2 subjects violated dominance in more than 10% of catch trials in experiment L2.f and 1 in experiment L2.h. These subjects were excluded from further analysis. In Experiment L2.e we decided in advance to take the conservative approach of removing people in the 5% of fastest or slowest people, all multiple submissions from the same IP address, and the 5% of people who alternated the most or the least between left and right responses. We removed 56 participants based on the above criteria that are not included on the reported sample size.

¥ In Experiments L2.a, L2.c and L2.d we ran fewer trials to make the duration of the experiment shorter. This allowed us to run the experiment with as many participants as possible within a fixed budget.

We also replicate SRH 1A, comparing positive-skew and negative-skew in the distributions of amounts in a different lab than SRH 1A (our Experiment L2.f). In Experiment L2.g we pushed the boundaries further by creating a distribution different from all the experiments reported in SRH and by collecting the data online. Finally, L2.h is a replication of SRH 2B and hence in the probability domain.

Figure 4: Revealed functions from the replication experiments in Level 2.



Note: Error bars are 95% confidence intervals.

To estimate the utility and probability weighting functions, we followed the same procedure as in Level 1; that is, we fitted Equations (2) and (3) to the choice data for the utility and the probability experiments respectively. Figure 4 shows the revealed utility and weighting functions from the set of Level 2 replications. For example, the top left panel of Figure 4 depicts the revealed utility functions from Experiment L2.a, where the resulting functions were more concave in the positive-skew treatment compared to the zero-skew treatment: Subjects assigned higher utilities to the common amounts of £100 and £200 when they experienced them in the positive-skew treatment (relative to the zero-skew treatment).

Eyeballing of Figure 4 reveals that the SRH effect is replicated in seven of the eight Level 2 experiments: in all cases apart from Experiment L2.e, we see a more concave utility function in the positive-skew treatment compared to the other (negative-skew or zero-skew) treatment. While we cannot rule out other interpretations, it seems possible that the failure to replicate in L2.e may be due to random error. Either way, however, this overall pattern of results is strong evidence that the SRH effect is replicable and robust to a range of changes in procedures and subject pools.

C. Level 3: Replication by implementing a new design

While our Level 2 analysis provides substantial evidence of robustness in the SRH effect, it does not show whether the conceptual interpretation of the original finding is correct. We therefore proceeded to a form of analysis in the spirit of Level 3 in Levitt and List's (2009) taxonomy. They suggest that Level 3 replication should involve creating alternative experimental designs, in order to test the same hypothesis as that tested in the original target of replication. We use a twofold strategy to stress test the explanation of the SRH effect based on DbS; first, we conduct new experiments in which, conditional on the truth of the DbS account, it may be harder for the SRH effect to work; second, we run tests in which, conditional on the truth of the DbS mechanism, we will eliminate the SRH effect altogether.

To this end, in one series of Level 3 experiments, which we refer to as *flagged* experiments, we used a within-subjects design and presented participants with choice options originating from two different choice sets, one with a positive-skew distribution and the other with a negative-skew distribution. For example, in Experiment L3.a, choices with prizes from, say, the positive-skew distribution were presented in videos by "Joanne", a young British woman. Choices with prizes from, say, the negative-skew distribution were presented by "Patrick", an older North American man with a significant beard. If participants can track the distributions of prizes separately depending on which speaker is making the offer, then DbS predicts that people will respond to the different speakers as if they have different utility functions for each speaker! In Experiment L3.b, the different-speaker manipulation was replaced with a different-product manipulation. In some choices people chose between

lotteries for holidays with different values and, in other choices, between lotteries for mobile phones with different values. Again, if people can track the values separately for the different types of product, DbS predicts people will respond as if they have different utility functions for different types of product value. Thus, if participants can simultaneously keep track of the two separate distributions and choose as if their decisions are informed by applying DbS separately to choices that come from the separate distributions, then retrospectively splitting the combined choice set into the original two sets for the analysis will yield the same effects as in the original between-subjects experiments. On the other hand, we do not know that individuals would try to use flags and separate the distributions accordingly. Even if they did, it seems possible that SRH effects might be reduced due to interference and imperfect memory. As such our flagged experiments provide a tougher environment for the operation of the mechanisms imputed by DbS.

Our second series of Level 3 replication experiments (which we label *non-flagged*) used the same within-subjects design just described, except we provided no flags (such as speaker or product type). Hence, in non-flagged experiments participants had no way of knowing what distribution the choices belonged to and therefore had no way of attributing choices to one distribution or another. As such, while a DbS model would still imply that any measured utility and probability weighting functions would depend on the background distributions of probabilities and amounts, in these experiments there is effectively only a single background distribution provided in the experiment. Yet, as experimenters, for the purpose of analysis we can still retrospectively split the data and analyse them separately for the two different distributions we used to generate the choices. Were we to do this, however, according to DbS, the SRH effect should disappear. So, if we continue to observe the SRH effect in our non-flagged designs that would be evidence against DbS's proposed mechanism as the cause of this context effect. Indeed, it would be a finding that no existing model we are aware of could account for.

Because of the completely within-subjects nature of these experiments, we reverted to the original SRH modelling by fitting one model to both conditions in an experiment. Random effects can now be estimated within one model, because all participants experience the attribute values from both distributions. In the logistic regression form, the model looks as follows for the experiments estimating a utility function:

$$\log\left[\frac{Prob(Choose\ safe)}{1-Prob(Choose\ safe)}\right] = \nu + \tau cond + \omega \log\left(\frac{q}{p}\right) + \xi cond \log\left(\frac{q}{p}\right) + \sum_i \beta_i X_i \quad (4)$$

An addition to the model described in Equation 2 are the terms involving $cond$. In expression 4 , $cond$ is a dummy to account for the two conditions (0 indicates the positive-skew; 1 indicates the other condition). Setting $\log(bias_{cond}) = \nu + \tau\ cond$, $\gamma_{cond} = \omega + \xi\ cond$,

and $u_{cond}(amount_i = \omega + exp(\beta_i/\gamma_{cond})$ follows, when Equation 1 is adapted to account for both conditions. To estimate probability weighting functions, we fit the following model:

$$\log\left[\frac{Prob(safe)}{1-Prob(safe)}\right] = \nu + \tau cond + \omega \log\left(\frac{y}{x}\right) + \xi cond \log\left(\frac{y}{x}\right) + \sum_i \beta_i X_i \quad (5)$$

and calculate the weights via $w_{cond}(p_i) = exp(\beta_i/\gamma_{cond})$. These are exactly the models estimated by SRH (see their Appendix A) and are extensions of our Equations 2 and 3.

Table 3 summarises the set of Level 3 experiments. As in Level 2, we varied the domain, the location, the incentives, the number of trials, the number of participants, and whether the experiment was conducted online or in the lab. For example, the second row of Table 3 describes the details of Experiment L3.a, which used the positive- vs. negative-skew in the distribution of amounts from the original Experiment SRH 1A and flagged them by labelling prizes as either mobile phones or holidays.

Table 3. Outline of the properties of all experiments from Level 3.

Replication (Original)	Skew	Domain	Amounts (£ or $)[±]	Probabilities (%)	Location (Sample)	N[≡]	No. Trials[⊥]	Incentives
Flagged Experiments[∅]								
L3.a (SRH 1A)	Positive vs. Negative	Utility	10, 20, 50,100, 200, 500 vs. 10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 2 (Student sample)	48	330	Course credit (non-incentivized)
L3.b (SRH 1A)	Positive vs. Negative	Utility	10, 20, 50,100, 200, 500 vs. 10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (Student sample)	42	160	Some for course credit some were paid £6 plus up to £5
Non-flagged Experiments								
L3.c (SRH 1A)	Positive vs. Negative	Utility	10, 20, 50, 100, 200, 500 vs. 10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 2 (Student sample)	45	450	Course credit (non-incentivized)
L3.d (SRH 1A)	Positive vs. Negative	Utility	10, 20, 50, 100, 200, 500 vs. 10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 2 (Student Sample)	50	450	£0 to £5
L3.e (SRH 1C)	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (50% MTurk and 50% Prolific Academic)	89	140	$2.25 or £1.50 and up to $/£25
L3.f (SRH 2B)	Positive vs. Negative	Probability	100, 200, 300, 400, 500 vs. 100, 200, 300, 400, 500	1, 2, 5, 10, 50, 99 vs. 1, 50, 90, 95, 98, 99	University 2 (Student sample)	49	390	£0 to £5

Notes: ±We used $ for all Amazon Mechanical Turk (MTurk) samples and £ for all Prolific Academic and student samples. In Experiments L3.d and L3.f, where the experiment was conducted using z-tree (Fischbacher, 2007) subjects were recruited via the online system ORSEE (Greiner, 2015), a show-up fee of £7 was added to their earnings.

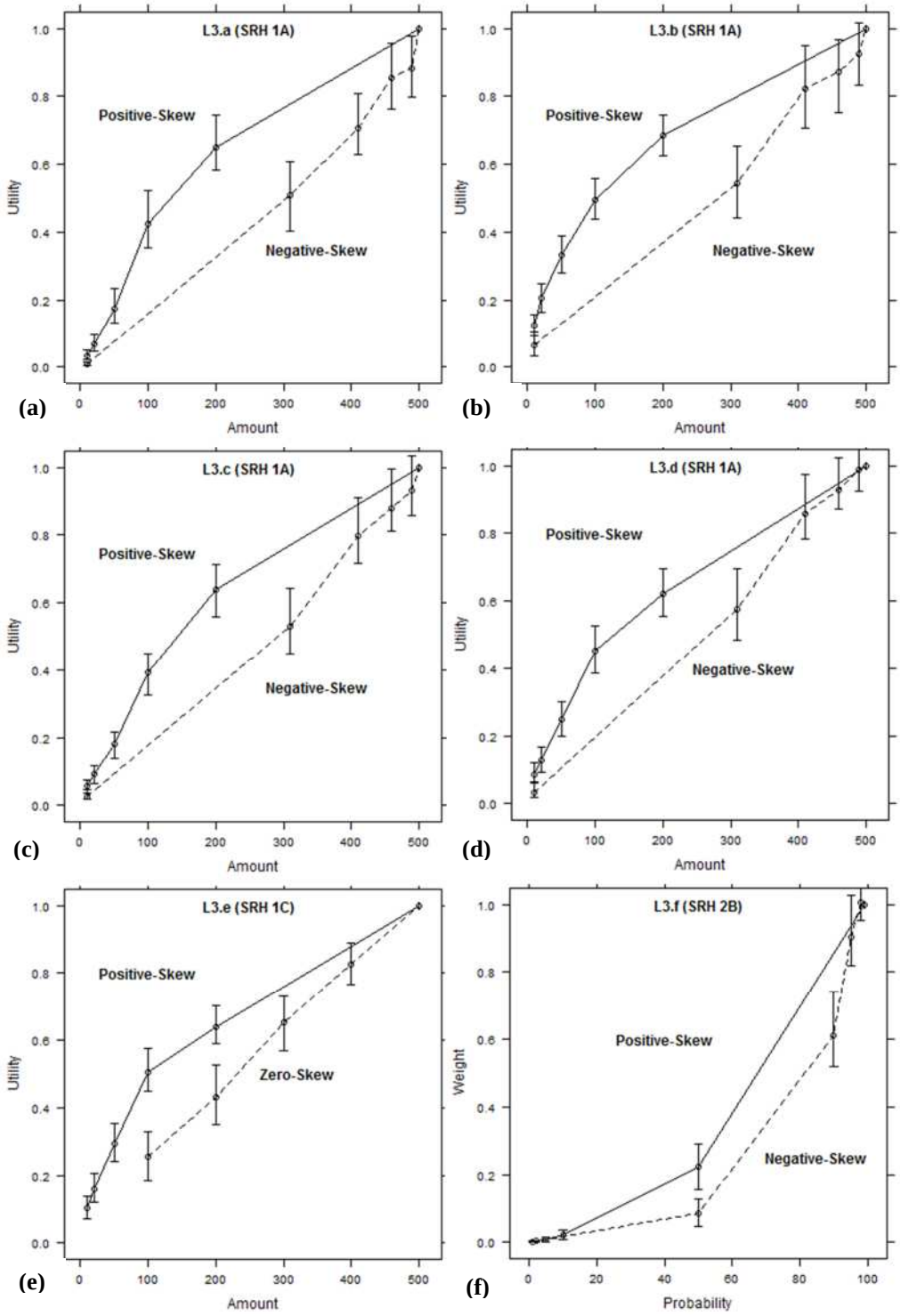
≡Catch trials were not included in replication Experiments L3.b and L3.e. 1 subject violated dominance in more than 10% of catch trials in Experiment L3.d and 3 in Experiment L3.f. These subjects were excluded from further analysis. In Experiment L3.e, we decided in advance to remove the 5% of fastest or slowest people, all multiple submissions from the same IP address, and the 5% of people who alternated the most or the least between left and right responses. We removed 32 participants on these criteria that are not included on the reported sample size.

⊥In Experiments L3.b, L3.e we ran fewer trials with as many participants as possible within a fixed budget.

∅The flagged experiments were as follows: In Experiment L3.a for half of the participants 150 positively skewed choices were framed as holidays and 150 negatively skewed choices were framed as mobile phones. For other participants it was the other way around; in Experiment L3.b the choice options from the different distributions were described by different people using video recordings. For half of the participants, Patrick described 80 gambles from the positively skewed set and Joanne described 80 gambles from the negatively skewed set; for the other half, this assignment was reversed.

We present the main results of Level 3 analysis in Figure 5.

Figure 5: Revealed functions from the replications of SRH in Level 3 using a within-subjects design.



Notes: L3.a-L3.b involve flagged choices, L3.c-L3.f do not. Error bars are 95% confidence intervals.

The top two panels of Figure 5 show results for the flagged experiments (L3.a and L3.b) where participants could potentially track what distribution the attributes in the choice set belonged to. The difference between the revealed utility functions for the two differently

distributed samples of amounts remain: the utility function using the choices from the choice set with positive-skew is more concave than the one from the choice set with negative-skew.

Surprisingly – and in contrast to DbS predictions – when estimating the revealed utility and weighting functions in the non-flagged experiments (L3.c-L3.f), the comparison of curves within each panel still shows the pattern of an SRH effect. It is hard to see how these differences can be rationalized with the DbS model, hence, the Level 3 results at least partially challenge the DbS interpretation of the SRH effect. We consider possible explanations for these results in Section IV below.

D. Level 4: Meta-analysis

The nineteen experiments we have analysed in this study (including the five Level 1 cases from the original SRH study) provide a large data set, based on the decisions of 1880 subjects, and are suitable for conducting a meta-analysis in order to examine the overall size, variability, and moderators of the effect of context on utility and probability weighting functions. We see this as a useful complement to the replication analysis of Levels 1 – 3 and so as a natural extension of the Levitt and List methodology, in cases where a suitable, comparably rich, data set has been generated.[6]

We first estimated the overall effect size using the differences in the revealed utility and weighting functions between conditions across all experiments for one attribute value. Subsequently, we calculated the effect sizes separately for: (i) the between-subjects design experiments; (ii) the within-subjects flagged experiments; and (iii) the within-subjects non-flagged experiments. From an explanatory point of view, the comparison of results for experiments in categories (i) and (iii) is especially interesting: If the within-subjects experiments' non-flagged effect size were as big as the between-subjects experiments' effect size, we could reject the DbS interpretation of the context effects, because the effect in the non-flagged experiments cannot emerge through the mechanisms proposed by DbS. However, if the effect size in the non-flagged within-subjects design experiments is smaller, DbS could still remain a candidate model for the interpretation of some part of the SRH effect.

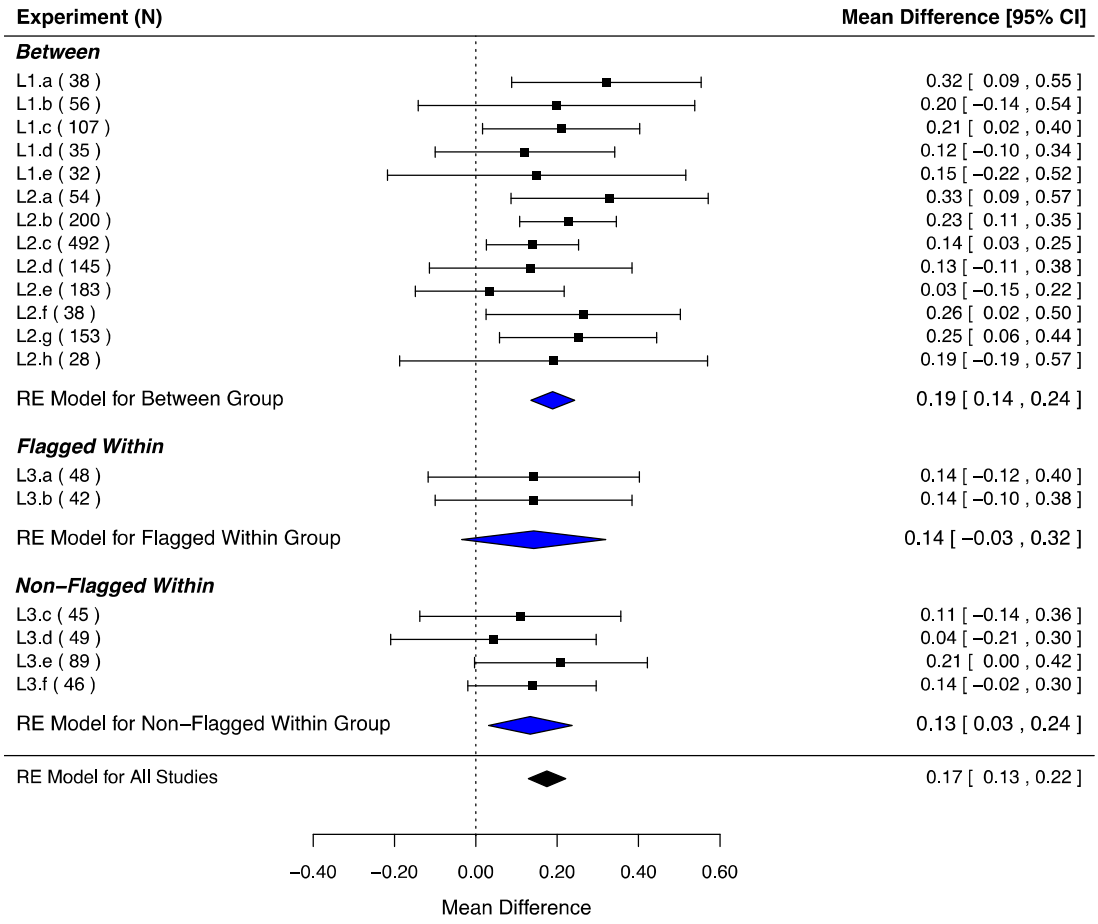
For the effect size measure, we identified the amount or the probability that was common to the two distributions in a given experiment and calculated the difference between the utility (or weighting) functions of the two distributions at that point. If no attribute was common within an experiment, we picked the two attribute values that were most similar to each other across the positive- and zero- or negative-skew distributions. For the experiments comparing zero- vs positive-skew, £200 (or $200) occurs in both distributions, so we

[6] We do not consider a meta-analysis of replication studies generated from a quasi-adversarial collaboration as a sufficient condition to obtain a definitive estimate of the underlying effect sizes. However, we regard it as an important step in the right direction.

subtracted the utility estimate of £200 in the zero-skew treatment from the utility of £200 in the positive-skew treatment. For experiments with positive- vs. negative-skew distributions of amounts, we calculated the difference between the estimated utility of £200 and the utility of £310 for the utility experiments (we used the £200 vs. £300 comparison for Experiment L2.g). Note that DbS predicts £200 in the positive-skew condition to be a higher utility than £310 in the negative-skew condition. The fact that £310 is a higher number than £200 makes this a conservative comparison. For the experiments manipulating the probability distribution we calculated the difference between the estimated probability weights of the common 30% (for SRH 2A replications) and the common 50% (for SRH 2B replications) in each condition.

We fitted a linear random effects model to estimate the effect of experimental design (between vs. flagged within vs. non-flagged within) and the skewness comparison (positive-negative vs. positive-zero) using mean differences.[7] The results of the meta-analysis are depicted in Figure 6.

Figure 6. Meta-analysis results from Level 4 replication.



Note: Mean Differences and 95% confidence intervals are shown as a function of the experimental design for between-subjects, and for flagged and non-flagged within-subject experiments.

[7] The estimations were obtained by using the metafor package in R by Viechtbauer (2010).

Figure 6 shows that our best estimate of the difference between the revealed utility or probability weight of the common attribute between the two distributions (positive-negative skew comparison or positive-zero skew comparison) overall is 0.17 95% CI [0.13, 0.22] on a scale where, arbitrarily, the utility of £500 for the utility experiments and the weight of 99% in the probability experiments are fixed at 1. Thus, taking all our experiments and SRH's experiments into consideration, the meta-analysis confirms a positive SRH effect in our data.

We estimated the effect of distribution comparison (positive-negative vs. positive-zero). The estimates do not differ between distribution comparisons, but could potentially differ by approximately 0.10 in each direction, ($beta_{Distribution}$ = 0.00 95% CI [-0.10, 0.09]). That is, the difference between the estimated utility of £200 from the positive-skew condition and £200 from the zero-skew condition is similar to the difference between the utility of £200 in the positive-skew condition and the utility of £310 in the negative-skew condition. However, there might be slight differences as the confidence interval is reasonably wide, leaving open the possibility that there are real differences between positive- vs zero-skew comparisons and positive- vs negative-skew comparisons.

Looking at the effect sizes for the three designs separately, the effect is largest in the between-subjects experiments ($MD_{between}$=0.19 95% CI [0.14, 0.24]). The effect is reduced in the flagged within-subjects experiments ($MD_{flagged\_within}$=0.14 95% CI [-0.03, 0.32]) and is smallest–with a reduction of 30% of the between-subjects experiments–in the non-flagged within-subjects experiments ($MD_{non\text{-}flagged\_within}$=0.13 95% CI [0.03, 0.24]). According to the meta-analytic model, the differences could be very small, or they could be opposite, such that the effects are larger in the non-flagged experiments, or it could be that the difference between the between-subjects experiments and the non-flagged experiments is about as large as the effect itself in the non-flagged experiments, ($beta_{Design}$ = 0.04 95% CI [-0.04, 0.12]). Given that the DbS account cannot apply to the non-flagged experiments, and the effect size in the non-flagged experiments is estimated to be similar to the two other groups, it is likely that much of the effect in the flagged and between-subject experiments should not be attributed to the DbS model. In light of this, what might explain the SRH effect in our data? The next section addresses this question directly.

III. Understanding the SRH effect

In the light of Level 3 non-flagged experiments showing an SRH effect even where it is not predicted by DbS, an obvious question to ask is whether the SRH effect might be explained as an artefact of this genre of experiment and its method of analysis. To motivate this possibility, we note that the recipe used to generate the choice sets in SRH style experiments, while based on a simple and coherent strategy for testing predictions of DbS,

may not generate sufficiently informative choice sets for unbiased parameter recovery, when applying the estimation procedures adopted by SRH and followed by us.[8]

The analysis is also characterised by a particular underlying choice model which may not be correctly specified. Therefore, it is appropriate to consider whether the SRH effect could be, at least in part, a systematic bias arising from some limitation of the model or estimation procedures to recover underlying preference parameters (if those exist) from the choice data these experiments generate.

We explore this possibility in two ways: first by seeing whether it is possible to identify patterns predicted by DbS which are akin to the SRH effect, but based on analysis that does not involve filtering the data through an estimating model; second, we examine how the SRH effect might emerge as an artefact of modelling inferences.

A. Model-free tests for effects predicted by Decision by Sampling

The data available to us provide a simple way of testing for an effect predicted by DbS that is closely akin to the SRH effect but using a form of analysis based on direct examination of choice data, thus short–circuiting the need for model fitting.

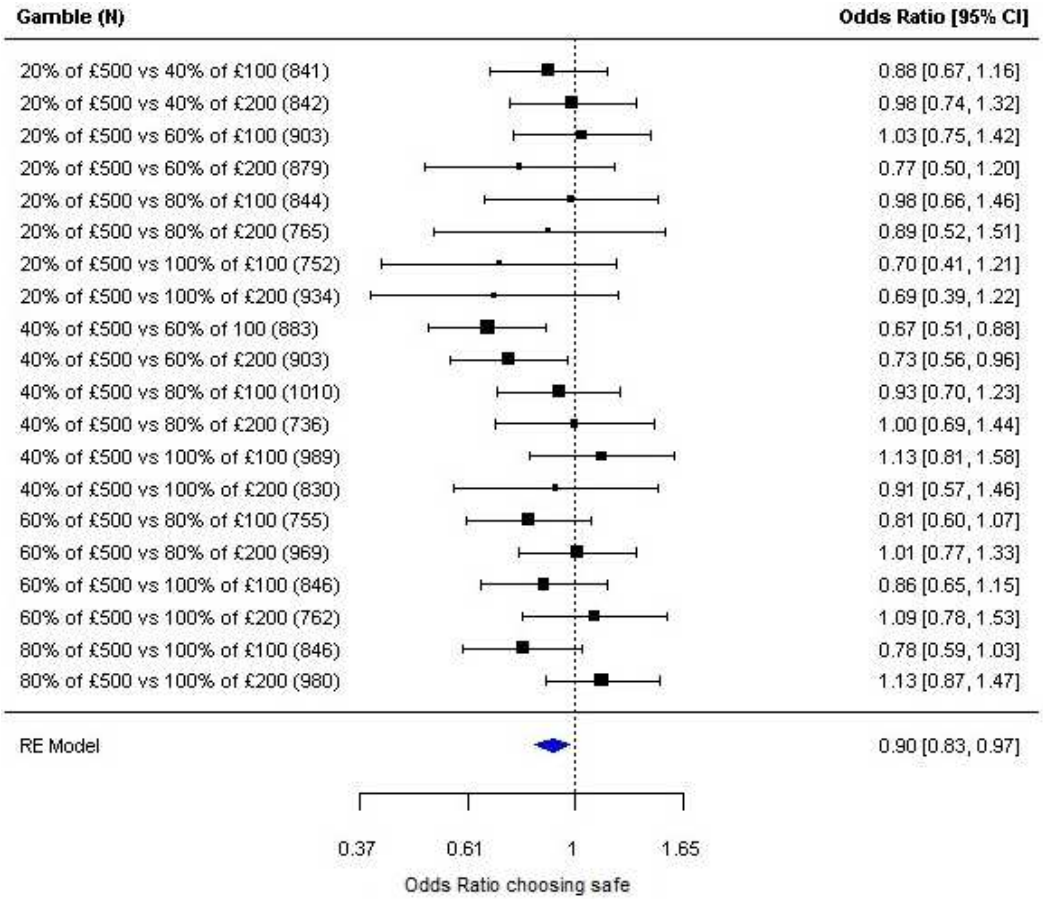
Within a given experiment, we can identify a number of *identical choices* for a given pair of gambles which occur in both treatments (or skews) of a given experiment. We focus this analysis on type 1C experiments because these provide numerous opportunities for this type of analysis and all of the comparisons are based on large sample sizes (see Figure 7 for details). Type 1C experiments feature money amounts of £10, £20, £50, £100, £200, and £500 in the positive-skew treatment and amounts of £100, £200, £300, £400, and £500 in the zero-skew treatment. Given the recipe for constructing choices, both treatments feature multiple choices between identical pairs of gambles each involving two of the three amounts £100, £200, £500. This affords a direct opportunity to test a prediction of DbS according to which the likelihood of choosing, say, the safer of the two gambles in any such pair should vary, predictably, between treatments. To see why, consider a choice between a 20% chance of £500 and a safer option of an 80% chance of £100. DbS predicts more frequent choice of the safer option in positive-skew treatments (where 100 is the third best outcome) as compared to when the same choice is embedded in the zero-skew (where £100 is the worst amount).

Within the data gathered for Type 1C experiments we identify 20 opportunities for testing whether such comparisons reveal support for predictions of DbS. We apply a random effects meta-analysis using odds ratios of the probability of choosing safe divided by the

[8] We are grateful to an anonymous referee for providing an analysis suggesting that differently skewed choice sets may indeed be differentially informative for revealing underlying preferences.

probability of choosing risky. Figure 7 shows the odds ratios for each of the 20 choices and their 95% confidence intervals. The number in brackets next to the description of each choice (e.g. 841 for the first row of the table) is the total number of decisions on which the relevant test is based, aggregating across the pair of treatments; while the number of subjects is not generally equal across a pair of treatments, every choice in every treatment involved at least 300 subjects. The overall odds ratio is 0.90, 95% CI [0.83, 0.97], which indicates that, across the set of 20 comparisons, subjects are less likely to choose the safer option when they are in the positive-skew as opposed to the zero-skew treatment.[9,10] The direction of this difference, however, is the opposite of that predicted by DbS. This mode of model-free analysis, therefore, provides no support for the effect predicted by DbS.

Figure 7. Meta-analysis of model-free tests (odds ratios and 95% confidence intervals).



[9] Odds ratios greater than 1 indicate an increase in the probability of choosing safe in the positive-skew treatment. Odds ratios less than one indicate a decrease.
[10] Our results also hold, qualitatively, using Pearson chi-square tests for pairwise comparisons.

B. Possible alternative explanations for the SRH effect

We now explore how the SRH effects observed in the model based analysis might have arisen, assuming they have nothing to do with DbS. We pursue this using simulation analysis. We begin by using a very simple preference model to generate synthetic choice data for the choice sets associated with both the 1A and 1C Experiments. We choose these two experiments because they are the ones which feature most commonly in the above replications. We then attempt to recover the preference function generating the data, using the SRH estimation procedure that we have applied throughout.

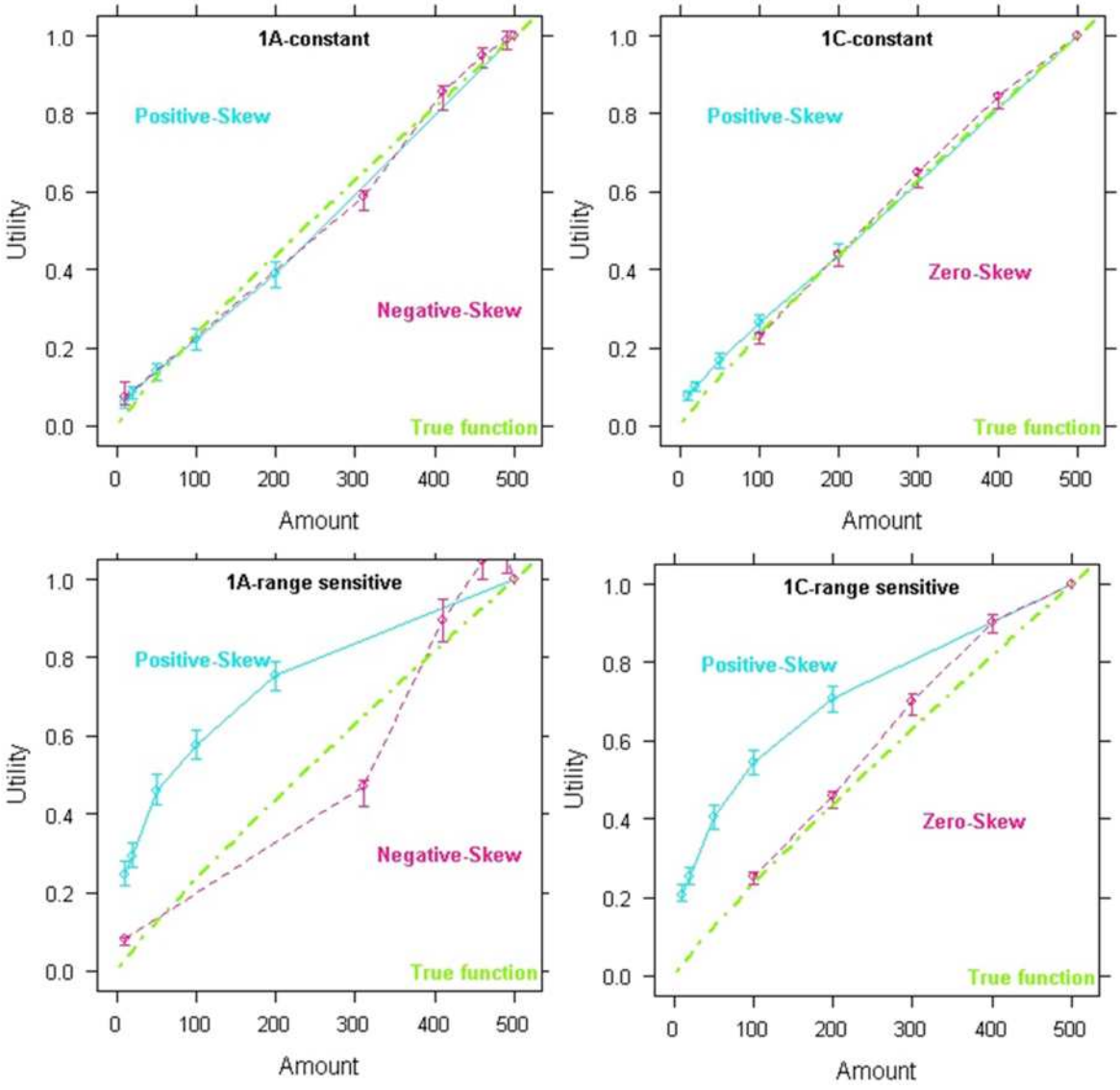
We generate data for 200 simulated agents who chose according to the sign of the *net perceived expected utility* = $[EU(C_1) - EU(C_2) + \varepsilon]$. In this expression, $EU(C_1)$ and $EU(C_2)$ are the expected utilities of a given pair of choice options where $\varepsilon \sim N(0, \sigma^2)$ is a random error, which may be interpreted as capturing imperfectly implemented preferences due to carelessness or miscalculation (see Hey and Orme, 1994). A simulated agent selects C_1 (resp. C_2) in any given choice, when the net perceived expected utility is positive (resp. negative). We use a standard power function for utility over consequences with $u(x)=x^{\alpha}$ and we repeat the exercise for three data sets generated using different, but unexceptional, values of α (i.e., 0.8, 0.9 and 1.0). Following standard approaches in the literature modelling real choice behaviour, we use two different error specifications with either *constant* variance; or *range sensitive* variance. In the first case, the error is drawn from the same distribution for each pairwise choice and σ^2 depends on the average range between payoffs in the choice set. In the second specification (following Bruhin et al. 2010; Fehr-Duda et al. 2010, 2011), the error variance is determined separately for each choice and is proportional to the outcome range for that choice. Both error specifications are widely used in studies that fit preference models to real choice data. Moreover, it has been shown that incorporating heteroscedastic error structures greatly improves the relative fit of expected utility (see Blavatskyy and Pogrebna, 2010; Buschena and Zilberman, 2000; Hey, 1995; Hey and Orme, 1994; Wilcox, 2011).

Figure 8 shows the results of simulations for Experiments 1A (left hand panels) and 1C (right hand panels). Results for all four panels are generated with $\alpha = 0.9$, but the qualitative patterns revealed in Figure 8 also hold for both the risk neutral and the more risk averse preferences too. The top two panels report results for the constant error model while the two bottom panels report results for the range sensitive error. In each panel, we plot separate functions estimated for choices based on positive and zero-skew choice sets. We also plot the true utility function used to generate the data.

Consider first the top two panels (constant error variance). At the eyeball level, the estimated functions approximate the true preferences rather well and there is no evidence of an SRH effect.

Things look quite different, however, when we consider the bottom panels (range sensitive error variance). In this case, we observe marked differences between the functions estimated on positive versus zero-skew choice sets; moreover, the differences we observe (i.e. more concave function for positive-skew) are consistent with the SRH effect predicted by DbS and those found in the experimental choice data.

Figure 8: Recovered functions for choices based on simulated data with added noise



Notes: Error bars are 95% confidence intervals.
Legend: Positive-Skew: Negative/Zero-Skew: True function:

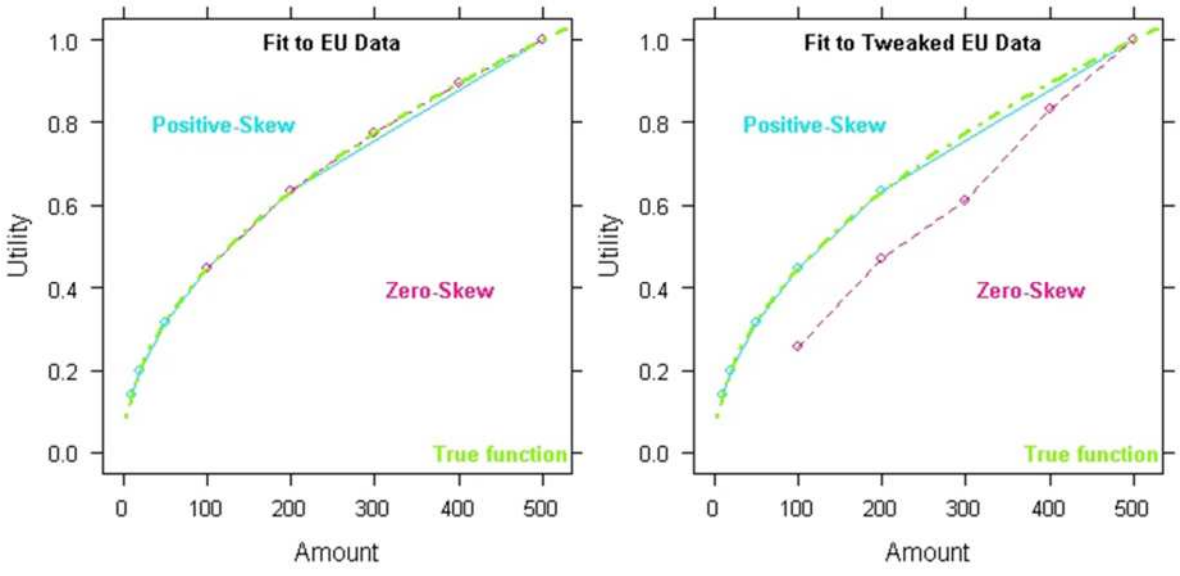
In this simulation exercise, we consistently find that systematic biases in the recovery of underlying preferences generate the SRH effect for agents who choose according to expected utility theory with an heteroskedastic error. While we do not wish to argue for the claim that expected utility preferences with range-sensitive errors provide a good general model of behaviour for real subjects in these experiments, we interpret this exercise as

demonstrating that failure to recover underlying preferences is a plausible candidate explanation for the SRH effect because we can mimic that effect in a simulated environment where agents choose according to a very simple and rather standard preference model.

Imposing a heteroskedastic error structure on the EU model is not the only way in which we can mimic the SRH effect in simulated data. To illustrate another possibility, consider an individual whose underlying preferences are represented by expected utility theory. Now imagine that same individual, but with some of their choices perturbed, relative to their underlying preferences, as a consequence of following a simple heuristic: specifically, *playing it safe when the stakes are high*. Evidence consistent with such an effect has been reported in various studies (e.g., Fehr-Duda et al. 2010; Holt and Laury, 2002, 2005; Lefebvre et al., 2010; Weber and Chapman, 2005). Here, we show that the operation of this heuristic is another candidate explanation of the SRH effect.

Following a similar process to that described above, we simulate choices from an EU model with power utility using choice sets having either positive or zero-skew. We then estimate preference functions, as before, using the approach we have adopted throughout. The left hand panel of Figure 9 presents revealed functions based on simulated choices for the choice sets corresponding with the 1C Experiments and with α=0.5. Notice that in this case, where the data generating process is a pure expected utility model, the fit is essentially perfect, regardless of whether it is estimated from choices simulated on the positive-skew or the zero-skew data.

Figure 9: Recovered functions for simulated choices: pure EU (left hand panel) and EU plus play it safe rule (right hand panel).



Note: Legend: Positive-Skew: Zero-Skew: True function:

The functions presented in the right hand panel of Figure 9 are generated in exactly the same way except that the simulated choices have been tweaked to incorporate an element of play it safe when the stakes are high. We do this by adding an increment to the log odds of safe choices for the small number of choices where the amounts available were both above 100 and only 100 apart: that is, for outcome pairs 200-300, 300-400 and 400-500. By design, this can only alter choices in the zero-skew condition and it has the effect of creating an SRH effect via the estimated function based on simulated choices for the zero-skew choice set becoming more linear.

Notice that the functions revealed for the tweaked data – i.e. with approximately linear function for the zero-skew treatment and markedly concave function for the positive-skew treatment – mimic closely the qualitative pattern revealed in the non-flagged Level 3 replication of Experiment 1C (i.e., Experiment L3.c shown in bottom left panel of Figure 5).

In line with the inference we drew from the simulation exercise appending the expected utility model with heteroskedastic error, we do not wish to argue for the claim that expected utility preferences combined with a play-it-safe heuristic constitute a good general account of our data and, in particular, the emergence of the SRH effect.[11] Our claim is more modest. We take these simulation exercises as an indication that the SRH effect may well be a consequence of model mis-specification. We have shown that if agents are approximately expected utility maximisers, but their choices depart from expected utility either as a consequence of a heteroskedastic error, or as a consequence of following a play-it-safe heuristic then, in either case, the SRH effect can be expected to emerge in functions estimated via the procedures used by SRH which take no account of these deviations from the preference model.

While we believe this analysis renders the SRH effect much less mysterious, we cannot, and do not, rule out other forms of mis-specification as contributory causes of the observed SRH effect (see Bhatia and Loomes, 2017, for further related discussion).

IV. Conclusion

We showed that the qualitative effects of attribute distributions on utility and probability weighting functions reported by Stewart, Reimers and Harris (2015) – which we have labelled the SRH effect - are highly replicable. We confirmed the analysis reported in Stewart, Reimers and Harris (2015) by reanalysing their original data using a slightly refined

[11] Using real choice data, Stewart et al. (2017) show that the residuals due to the playing-it-safe-when-the-stakes-are-high are alone sufficient to create the SRH effect. They explore the broader implications of this for the non-generalisation of utility functions estimated in different choice sets.

approach. We then replicated the SRH effect across a set of fresh experiments using designs approximating their original setup.

Notwithstanding our ability to replicate the SRH effect as just described, however, results reported above based on new experiments, model-free analysis of choice data and simulation analysis of synthetic data, together, cast the SRH effect in an entirely new light. A key result is that we continued to find the SRH effect in new experiments designed such that the Decision by Sampling model no longer predicts it. We also failed to find any evidence of effects predicted by Decision by Sampling in model-free analysis of the experimental choice data. Given these results, we explored alternative explanations for it. Using parameter recovery simulations, we were able to identify two candidate explanations for the SRH effect. First, we showed that misspecification of the stochastic form of EU can systematically bias the estimated utility and probability weighting functions in line with the SRH effect. Second, we showed that the SRH effect arises when a stochastic EU model is fitted to simulated data that has been tweaked to incorporate a simple decision heuristic (playing safe when stakes are high). Both mechanisms are plausible candidates for explaining the phenomenon reported by SRH.

While we do not interpret the SRH effect as evidence for Decision by Sampling, our analysis does not imply that the apparent instability of preference functions identified by the original SRH paper should be dismissed as irrelevant from the point of view of those seeking to model risk preferences, or to elicit them from choice behaviour. SRH interpreted their evidence as showing that the shape of the utility functions and weighting functions estimated from choice data are not just a property of the decision maker, but they are fundamentally context dependent and vary with features of the choice environment. While our analysis challenges the idea that the SRH effect is evidence of any irreducible context sensitivity, we accept that it is nevertheless evidence of a genuine issue, albeit a more familiar or traditional one to do with model specification. To the extent that our mis-specification interpretation of the SRH effect is correct, in principle, the SRH effect can be avoided by fitting the “right model” and thereby eliminating context dependence. In practice, however, the right model may be an elusive creature and hence problems of mis-specification and associated context dependency may often be difficult to avoid.

References

Bateman, I., Kahneman, D., Munro, A., Starmer, C., & Sugden, R. (2005). Testing competing models of loss aversion: An adversarial collaboration. *Journal of Public Economics*, *89*(8), 1561-1580.

Bhatia, S., & Loomes, G. (2017). Noisy preferences in risky choice: A cautionary note. *Psychological review*, *124*(5), 678.

Blavatskyy, P. R., & Pogrebna, G. (2010). Models of stochastic choice and decision theories: Why both are important for analyzing decisions. *Journal of Applied Econometrics*, *25*(6), 963-986.

Bruhin, A., Fehr-Duda, H., & Epper, T. (2010). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, *78*(4), 1375-1412.

Buschena, D., & Zilberman, D. (2000). Generalized expected utility, heteroscedastic error, and path dependence in risky choice. *Journal of Risk and Uncertainty*, *20*(1), 67-88.

Camerer et al. (2016), *Science*, 10.1126/science.aaf0918

Corrigan, J. R., Drichoutis, A. C., Lusk, J. L., Nayga, R. M., & Rousu, M. C. (2012). Repeated Rounds with Price Feedback in Experimental Auction Valuation: An Adversarial Collaboration. *American Journal of Agricultural Economics*, *94*(1), 97-115.

Drichoutis, A. C., & Nayga, R. M. (2013). Eliciting risk and time preferences under induced mood states. *The Journal of Socio-Economics*, *45*, 18-27.

Etchart-Vincent, N. (2004). Is probability weighting sensitive to the magnitude of consequences? An experimental investigation on losses. *Journal of Risk and Uncertainty*, *28*(3), 217-235.

Fehr-Duda, H., Bruhin, A., Epper, T., & Schubert, R. (2010). Rationality on the rise: Why relative risk aversion increases with stake size. *Journal of Risk and Uncertainty*, *40*(2), 147-180.

Fehr-Duda, H., Epper, T., Bruhin, A., & Schubert, R. (2011). Risk and rationality: The effects of mood and decision rules on probability weighting. *Journal of Economic Behavior & Organization*, *78*(1), 14-24.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, *10*(2), 171-178.

Friedman, D., Isaac, R. M., James, D., and Sunder, S. (2014). *Risky curves. On the empirical failure of expected utility*, Princeton University Press, Routledge, NY.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, (1) 1-12.

Hey, J. D. (1995). Experimental investigations of errors in decision making under risk. *European Economic Review*, *39*(3-4), 633-640.

Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica: Journal of the Econometric Society*, 1291-1326.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, *92*(5), 1644-1655.

Holt, C. A., & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review*, *95*(3), 902-912.

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, *58*(9), 723.

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the Erez–Latham dispute regarding participation in goal setting. *Journal of Applied Psychology*, *73*(4), 753.

Lefebvre, M., Vieider, F. M., & Villeval, M. C. (2010). Incentive effects on risk attitude in small probability prospects. *Economics Letters*, *109*(2), 115-120.

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: the past, the present, and the future. *European Economic Review*, *53*(1), 1-18.

Luce RD (1959) *Individual Choice Behavior* (John Wiley & Sons, New York).

McFadden, D. L. (1976). Quantal choice analysis: A survey. In *Annals of Economic and Social Measurement, Volume 5, number 4* (pp. 363-390). NBER.

McFadden, D. (2001). Economic choices. *The American Economic Review*, *91*(3), 351-378.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269-275.

Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *The American Economic Review*, *104*(1), 277-290.

Open Science Collaboration. (2015, 28 August). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Shepard RN (1957) Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrica* 22(4):325–345.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology*, *62*, 1041–1062. doi:10.1080/17470210902747112

Stewart, N., Canic, E., & Mullett, T., (2017). On the futility of estimating utility functions: Why the parameters we measure are wrong, and why they do not generalize. Manuscript submitted for publication.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*, 1-26. doi:10.1016/j.cogpsych.2005.10.003

Stewart, N., Reimers, S., & Harris, A. J. L. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. Management Science, 61, 687-705. doi: 10.1287/mnsc.2013.1853

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48.

Weber BJ, Chapman G.B. (2005). Playing for peanuts: Why is risk seeking more common for low-stakes gambles? Organ. Beh. Hum. Dec. 97:31–46, URL http://dx.doi.org/10.1016/j.obhdp.2005.03. 001.

Wedell, D. H. (1991). Distinguishing among models of contextually induced preference reversals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 767.

Appendix A

The calculation errors in SRH's original analysis that were discovered as a result of our Level 1 replication analysis are as follows: In Experiment SRH 1A the original analysis takes into account only 120 instead of 150 questions. However, except from moving the functions slightly, this has no effect on the difference between the two conditions. In the description of the analysis of Experiment SRH 2B it is mentioned that none of the participants violated dominance in more than 10% of the trials. Our analysis shows that 4 participants violated dominance and therefore should have been excluded from further analysis. Again, the differences between conditions reported by SRH remain significant after this adjustment.

As we explain in the main paper, the Level 1 analysis we apply and report in the manuscript differs from SRH's original analysis in terms of the estimation of both the confidence intervals (i.e., we used a more reliable bootstrapping method) and the revealed functions (i.e., we estimated the revealed functions separately for each condition). To further probe the robustness of the inference from SRH's as well as our own (Level 1) analysis, it is useful to test statistically whether SRH conclusions would still hold when correcting the above calculation errors. Hence, we applied SRH's original analysis (without the refinements we use in the analysis reported in the main paper) to SRH raw data and compared the results with and without taking into account these errors. The results from this comparison are reported in Table A1 below. Even though there are some minor differences in the reported statistics with regard to Experiments SRH 1A and SRH 2B, the results are qualitatively very similar and none of SRH's original claims are affected by correction of these minor errors.

Table A1. Results from the SRH's original analysis using SRH raw data with and without the calculation errors.

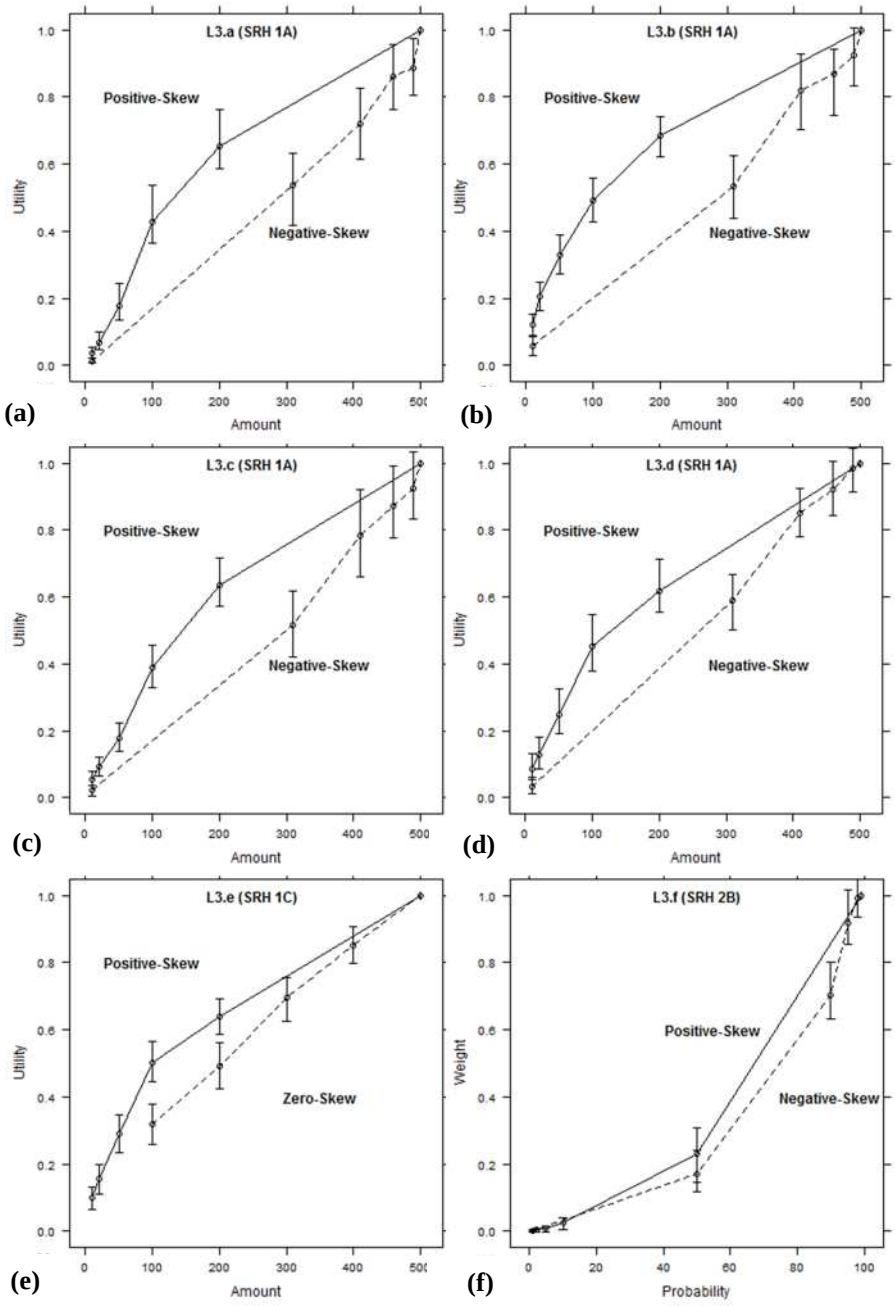
	With the calculation errors		Without the calculation errors	
Experiment	Differences in concavity in the revealed functions	Differences in weighting common amounts or probabilities	Differences in concavity in the revealed functions	Differences in weighting common amounts or probabilities
SRH 1A	$\chi^2(1)=6.36$, $p=0.012$	$\chi^2(1)=7.05$, $p=0.0079$ between £200 and £310	$\chi^2(1)= 11.93$, $p=.0006$	$\chi^2(1)=22.75$, $p<0.0001$ between £200 and £310
SRH 1B	$\chi^2(1)=6.99$, $p=.0082$	$\chi^2(1)=26.96$, $p<0.0001$ for the common £100, $\chi^2(1)=7.16$, $p=0.0074$ for the common £200	$\chi^2(1)=6.98$, $p=.0082$	$\chi^2(1)=27.20$, $p<0.0001$ for the common £100, $\chi^2(1)=6.77$, $p=.009$ for the common £200
SRH 1C	$\chi^2(1)=3.50$, $p=0.06$	$\chi^2(1)=59.79$, $p<0.0001$ for the common £100, $\chi^2(1)=50.47$, $p<0.0001$ for the common £200	$\chi^2(1)=3.49$, $p=.06$	$\chi^2(1)=59.75$, $p<0.0001$ for the common £100, $\chi^2(1)=50.32$, $p<0.0001$ for the common £200
SRH 2A	$\chi^2(1)=2.18$, $p=0.13$	$\chi^2(1)=18.18$, $p<0.0001$ for the common 30%, $\chi^2(1)=14.31$, $p=0.0002$ for the common 70%	$\chi^2(1)=2.18$, $p=0.14$	$\chi^2(1)=18.22$, $p<0.0001$ for the common 30%, $\chi^2(1)=14.41$, $p=0.0001$ for the common 70%
SRH 2B	$\chi^2(1)=181.5$, $p<0.0001$	$\chi^2(1)=41.72$, $p<0.0001$ for the common 50%	$\chi^2(1)=119.6$, $p<0.0001$	$\chi^2(1)=22.8$, $p<0.0001$ for the common 50%

Appendix B

In our Level 3 replication in order to account for the within-subject nature of this series of experiments, we used only one model including all random effects, instead of using separate models for each condition as we did for the between-subject experiments.

Figure B1 shows the functional forms and the relative confidence intervals obtained by estimating a separate model for each condition. By comparing Figure B1 to Figure 5 in the manuscript, it is possible to see that both the estimated functional forms and the confidence intervals are very similar across the two figures.

Figure B1: Revealed functions from the replications of SRH in Level 3 using 2 models instead of 1.



Note: Error bars are 95% confidence intervals.